Lab Session 2

Some UNIX programs

```
wc -l # count lines
LC_CTYPE=C printf '%d' "'A" # get the ASCII value of A
```

SQL*Loader

```
-- table schema:
CREATE TABLE codesPostaux (
insee varchar2(6),
nom_commune varchar2(50),
zip varchar2(6),
LIBELLE varchar2(50),
dum1 varchar2(50)
);

-- control file control.txt:
LOAD DATA INFILE 'codes_postaux.csv'
TRUNCATE
INTO TABLE codesPostaux
FIELDS TERMINATED BY ';'
( insee ,
nom_commune,
zip,
libelle,
dum1
)

-- script: (change xxx for your login, of course)
sqlldr userid=C##xxxx_a/xxxx_a control=control.txt log=log.txt bad=bad.txt  \
direct=y errors=0 skip=1
```

Quelques liens utiles:

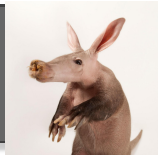
http://www.oracle-dba-online.com/sql_loader.htm
https://docs.oracle.com/database/121/SUTIL/GUID-8D037494-07FA-4226-B507-E1B2ED10C144.htm

Regular Expressions:

Lab. Ex 1.1 (egrep)

1. Use egrep to list words beginning with "aa" in the `/usr/share/dict/words` dictionnary.
2. Count words containing the substring "hard" in this dictionnary.
3. List words having a 6-letter substring none of which are vowels. To keep things simple you may consider accented vowels to be non-vowels. But you must avoid punctuation symbols. For an optimal solution consider using equivalence classes.
4. Do some words have a letter repeated three times in a row?

```
3 words begin with 'aa'
84 words contain 'hard'
3 words have a substring with 6 consonants, 27 if we include accented vowels.
2 words have a triples letter
```



Lab. Ex 1.2 (A special character class)

Which pattern do you think that regular expression "b[--a]b" will match? Check it!

Lab. Ex 1.3 (Oracle SQL)

1. Load the file codes_postaux.csv in a table, using SQL*Loader
2. Select in SQL the postal code and name of cities whose name contains the substring VIGNOBLE.
3. Count city name referencing a saint. To simplify, we will apply following rules: we count "saint" or "st" as a distinct word, such as in "st Eloi", or "bourg saint cristophe", but we will discard forms such as "tressaint", "Saint-Arnac" , and won't care about women saints nor altered forms such as "Sanary".
4. Update all INSEE codes of the form "2A..." into "20...".

Lab. Ex 1.4 (Python)

1. Edit file `codes_postaux.csv` to replace INSEE codes of the form "2A..." into "20..." (easier with `sed` than with `python`).
2. Write a python script taking as input a string s and file f, and returns the list of words following s in f.

   You may test your code on files `macbeth.txt` et `rj.txt` with strings "Thou art" and "As pretty as", or whatever else comes to your mind.